Noname manuscript No.
(will be inserted by the editor)

Variable Selection for Model-Based Clustering using the Integrated Complete-Data Likelihood

Matthieu Marbac · Mohammed Sedki

Received: date / Accepted: date

Abstract Variable selection in cluster analysis is important yet challenging. It can be achieved by regularization methods, which realize a trade-off between the clustering accuracy and the number of selected variables by using a lasso-type penalty. However, the calibration of the penalty term can suffer from criticisms. Model selection methods are an efficient alternative, yet they require a difficult optimization of an information criterion which involves combinatorial problems. First, most of these optimization algorithms are based on a suboptimal procedure (*e.g.* stepwise method). Second, the algorithms are often greedy because they need multiple calls of EM algorithms. Here we propose to use a new information criterion based on the integrated complete-data likelihood. It does not require the maximum likelihood estimate and its maximization appears to be simple and computationally efficient. The original contribution of our approach is to perform the model selection without requiring any parameter estimation. Then, parameter inference is needed only for the unique selected model. This approach is used for the variable selection of a Gaussian mixture model with conditional independence assumption. The numerical experiments on simulated and benchmark datasets show that the proposed method often outperforms two classical approaches for variable selection. The proposed approach is implemented in the R package VarSelLCM available on CRAN.

Keywords Gaussian mixture model · Information criterion · Integrated complete-data likelihood · Model-based clustering · Variable selection

1 Introduction

Clustering allows us to summarize large datasets by grouping individuals into few characteristic classes. It aims to discover an *a priori* unknown partition among the individuals. In many cases, this partition may be best explained by only a subset of the observed variables. So, by performing the variable selection in the cluster analysis, both *model fitting* and *result interpretation* are facilitated. Indeed, for a fixed sample size, a variable selection method can provide a more accurate identification of the classes. Moreover, such methods bring out the variables characterizing the classes .

Regularization methods can be used to achieve variable selection in clustering. One can cite the approaches of Friedman, J.H. and Meulman, J.J. (2004) or Pan, W. and Shen, X. (2007). Recently, these methods have been outperformed by the *sparse K-means* proposed by Witten and Tibshirani (2010). It uses a lasso-type penalty to select the set of variables relevant to clustering. Since it requires small computational times, it can manage high-dimensional datasets. Moreover, the selection of the number of classes is a difficult issue since probabilistic tools are not available. Finally, its results are sensitive to structure of the penalty term.

M. Marbac
INSERM U1181
E-mail: matthieu.marbac@inserm.fr

M. Sedki
INSERM U1181 and University of Paris Sud
E-mail: mohammed.sedki@inserm.fr

Model selection approaches can be used to carry out the variable selection in a probabilistic framework. Tadesse, M.G. and Sha, N. and Vannucci, M. (2005) consider two types of variables: the set of the *relevant variables* and the set of the *irrelevant variables* which are independent of the relevant ones. This method has been extended by Raftery and Dean (2006) by using a greedy search algorithm to find the set of relevant variables. Obviously, this algorithm finds only a local optimum in the space of models. It is feasible for quite large datasets because of its moderate computing time. Still, this method remains time consuming since the model comparisons are performed by using the BIC criterion (Schwarz, 1978). Therefore, the maximum likelihood estimate must be computed for each competing model. These estimates are mainly instrumental since the practitioner interprets only the estimate related to the best model.

In this paper, we propose a new information criterion, named *MICL criterion* (Maximum Integrated Complete-data Likelihood), for carrying out the variable selection in model-based clustering. This criterion is quite similar to the ICL criterion (Biernacki, C. and Celeux, G. and Govaert, G., 2010), and it inherits its main properties. However, these two criteria evaluate the integrated complete-data likelihood at two different partitions. The MICL criterion uses the partition maximizing this function, while the ICL criterion uses the partition provided by a MAP rule associated to the maximum likelihood estimate.

In this article, we focus on variable selection for a *Gaussian mixture model with conditional independence* assumption, but the method can be extended to more general mixture models. Note that this model is useful especially when the number of variables is large (Hand and Keming, 2001). Moreover, the conditional independence is often an hidden assumption made by distance-based methods like the *K-means* algorithm (Govaert, 2009). The MICL criterion takes advantage of the closed form of the integrated complete-data likelihood when the priors are conjugated. The model selection is carried out by a simple and fast procedure which alternates two maximizations for providing the model maximizing the MICL criterion. The convergence properties of this algorithm are similar to the convergence properties of the EM algorithm. In particular, it converges to a local optimum of the function to maximize. So, multiple random initializations are required to ensure its convergence to the global maximum.

The proposed method and the methods of Witten and Tibshirani (2010), and of Raftery and Dean (2006) are compared on simulated and on challenging real datasets. We show that the proposed method outperforms both other methods in terms of model selection and partitioning accuracy. It often provides a model with a better value of the BIC criterion than the algorithm of Raftery and Dean (2006), although it does not directly optimize this criterion. Finally, we show that the proposed method can manage datasets with a large number of variables and a moderately large number of individuals. Note that it is the most common situation which requires variable selection in cluster analysis.

The paper is organized as follows. Section 2 briefly reviews the framework of variable selection for the Gaussian mixture model. A presentation of the integrated complete-data likelihood is done in Section 3 before introducing the MICL criterion. Section 4 is devoted to the inference based on the MICL criterion. Section 5 illustrates the robustness properties of the MICL criterion and compares the three methods of variable selection on simulated data. Section 6 compares the three methods of variable selection on challenging datasets. The advantages and limitations of the method are discussed in Section 7.

2 Variable selection for Gaussian mixture model

2.1 Mixture model of Gaussian distributions

Data to analyze are n observations $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, where object $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$ is described by d continuous variables defined on $\mathbb{R}^d$. Observations are assumed to arise independently from a Gaussian mixture model with g components, assuming conditional independence between variables. Therefore, the model density is written as

$$f(\boldsymbol{x}_i|\boldsymbol{m}, \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \prod_{j=1}^{d} \phi(x_{ij}|\mu_{kj}, \sigma_{kj}^2), \tag{1}$$

where $\boldsymbol{m}$ specifies the model, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})$ is the whole parameter vector, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ is the vector of mixing proportion defined on the simplex of size g, $\boldsymbol{\mu} = (\mu_{kj}; k = 1, \ldots, g; j = 1, \ldots, d)$, $\boldsymbol{\sigma} = (\sigma_{kj}; k = 1, \ldots, g; j = 1, \ldots, d)$, and where $\phi(.|\mu_{kj}, \sigma_{kj}^2)$ is the density of a univariate Gaussian distribution with mean μ_{kj} and variance σ_{kj}^2.

A variable is said to be *irrelevant* to the clustering if its one-dimensional marginal distributions are equal between components. Thus, by introducing ω_j such that $\omega_j = 0$ if variable j is irrelevant and $\omega_j = 1$ if the variable is *relevant* for the clustering, the following equalities hold:

$$\forall j \in \{j' : \omega_{j'} = 0\},\ \mu_{1j} = \ldots = \mu_{gj} \text{ and } \sigma_{1j} = \ldots = \sigma_{gj}. \tag{2}$$

Thus, a model $\boldsymbol{m} = (g, \boldsymbol{\omega})$ is defined by a number of components g and the binary vector $\boldsymbol{\omega} = (\omega_j; j = 1, \ldots, d)$ which encodes whether each of d possible variables are relevant to the clustering.

2.2 Model selection based on the integrated likelihood

Model selection generally aims to find the model $\hat{\boldsymbol{m}}$ which obtains the highest posterior probability among a collection of competing models $\mathcal{M}$. So,

$$\hat{\boldsymbol{m}} = \text{argmax}_{\boldsymbol{m} \in \mathcal{M}} p(\boldsymbol{m}|\mathbf{x}). \tag{3}$$

This model selection approach is consistent since $\hat{\boldsymbol{m}}$ converges in probability to the true model $\boldsymbol{m}^{(0)}$ as long as the true model belongs to the model space (*i.e.* if $\boldsymbol{m}^{(0)} \in \mathcal{M}$).

By assuming uniformity for the prior distribution of $\boldsymbol{m}$, $\hat{\boldsymbol{m}}$ maximizes the integrated likelihood defined by

$$\begin{aligned} \hat{\boldsymbol{m}} &= \text{argmax}_{\boldsymbol{m} \in \mathcal{M}} p(\mathbf{x}|\boldsymbol{m}) \text{ with } p(\mathbf{x}|\boldsymbol{m}) \\ &= \int_{\boldsymbol{\Theta}_{\boldsymbol{m}}} p(\mathbf{x}|\boldsymbol{m}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{m}) d\boldsymbol{\theta}, \end{aligned} \tag{4}$$

where $\boldsymbol{\Theta}_{\boldsymbol{m}}$ is the parameter space of model $\boldsymbol{m}$, $p(\mathbf{x}|\boldsymbol{m}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{m}, \boldsymbol{\theta})$ is the likelihood function and $p(\boldsymbol{\theta}|\boldsymbol{m})$ is the prior distribution of the parameters. We assume independence between the prior, so

$$p(\boldsymbol{\theta}|\boldsymbol{m}) = p(\boldsymbol{\pi}|\boldsymbol{m}) \prod_{j=1}^{d} p(\boldsymbol{\sigma}^2_{\bullet j}, \boldsymbol{\mu}_{\bullet j}|\boldsymbol{m}), \tag{5}$$

where $\boldsymbol{\sigma}^2_{\bullet j} = (\sigma^2_{kj}; k = 1, \ldots, g)$ and $\boldsymbol{\mu}^2_{\bullet j} = (\mu^2_{kj}; k = 1, \ldots, g)$, and

$$p(\boldsymbol{\sigma}^2_{\bullet j}, \boldsymbol{\mu}_{\bullet j}|\boldsymbol{m}) = \Big(\prod_{k=1}^{g} p(\sigma^2_{kj}|\boldsymbol{m}) p(\mu_{kj}|\boldsymbol{m}, \sigma^2_{kj}) \Big)^{\omega_j} \Big(p(\sigma^2_{1j}|\boldsymbol{m}) p(\mu_{1j}|\boldsymbol{m}, \sigma^2_{1j}) \Big)^{1-\omega_j}.$$

We use conjugate prior distributions, thus $\boldsymbol{\pi}|\boldsymbol{m}$ follows a Dirichlet distribution $\mathcal{D}_g(\frac{1}{2}, \ldots, \frac{1}{2})$ which is the Jeffreys non informative prior (Robert, 2007). Moreover, $\sigma^2_{kj}|\boldsymbol{m}$ follows an Inverse-Gamma distribution $\mathcal{IG}(\alpha_j/2, \beta_j^2/2)$ and $\mu_{kj}|\boldsymbol{m}, \sigma^2_{kj}$ follows a Gaussian distribution $\mathcal{N}(\lambda_j, \sigma^2_{kj}/\delta_j)$, where $(\alpha_j, \beta_j, \lambda_j, \delta_j)$ are hyper-parameters.

Unfortunately, the integrated likelihood is intractable. However, many methods permit to approximate its value (Friel, N. and Wyse, J., 2012). The most popular approach consists in using the BIC criterion (Schwarz, 1978), which approximates the logarithm of the integrated likelihood by Laplace approximation and requires maximum likelihood estimation. The BIC criterion is written as

$$\text{BIC}(\boldsymbol{m}) = \ln p(\mathbf{x}|\boldsymbol{m}, \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) - \frac{\nu_{\boldsymbol{m}}}{2} \ln n, \tag{6}$$

where $\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}$ is the maximum likelihood estimate related to model $\boldsymbol{m}$ when $\nu_{\boldsymbol{m}}$ is the number of parameters required by $\boldsymbol{m}$.

For a fixed value of g, the variable selection in clustering necessitates the comparison of 2^d models. Therefore, an exhaustive approach which approximates the integrated likelihood for each competing model is not doable. Instead, Raftery and Dean (2006) carry out the model selection by deterministic algorithms (like a *forward* method) which are suboptimal. Moreover, they are time consuming when the number of variables is large, because they involve many parameter estimations for their model comparisons.

All maximum likelihood estimates are mainly instrumental: they are only used for computing the BIC criterion, with the exception of the estimates related to the selected model $\hat{\boldsymbol{m}}$ which are interpreted by the practitioner. Therefore, we introduce a new criterion for model selection which does not require parameter estimates.

3 Model selection based on the integrated complete-data likelihood

3.1 The integrated complete-data likelihood

A partition is given by the vector $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ig})$ indicates the class label of vector i, *i.e.* $z_{ik} = 1$ if $\boldsymbol{x}_i$ arises from component k and $z_{ik} = 0$ otherwise. In cluster analysis, $\mathbf{z}$ is a missing value. Thus, the likelihood function computed on the complete-data (observed and latent variables), called *complete-data likelihood* function, is introduced. It is defined by

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{g} \left(\pi_k \prod_{j=1}^{d} \phi(x_{ij}|\mu_{kj}, \sigma_{kj}^2)\right)^{z_{ik}}. \tag{7}$$

The *integrated complete-data likelihood* is

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}) = \int_{\boldsymbol{\Theta_m}} p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{m}) d\boldsymbol{\theta}. \tag{8}$$

Since conjugate prior distributions are used, the integrated complete-data likelihood has the following closed form

$$p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}) = p(\mathbf{z}|g) \prod_{j=1}^{d} p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z}), \tag{9}$$

where $\mathbf{x}_{\bullet j} = (x_{ij}; i = 1, \ldots, n)$. More specifically,

$$p(\mathbf{z}|g) = \frac{\Gamma(\frac{g}{2})}{\Gamma(\frac{1}{2})^g} \frac{\prod_{k=1}^{g} \Gamma(n_k + \frac{1}{2})}{\Gamma(n + \frac{g}{2})}, \tag{10}$$

where $n_k = \sum_{i=1}^{n} z_{ik}$, and

$$p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z}) = \begin{cases} \left(\frac{1}{\pi}\right)^{n/2} \frac{\Gamma\left(\frac{n+\alpha_j}{2}\right)}{\Gamma\left(\frac{\alpha_j}{2}\right)} \left(\frac{\beta_j^{\alpha_j}}{s_j^{\alpha_j+n}}\right) \sqrt{\frac{\delta_j}{n+\delta_j}} \text{ if } \omega_j = 0 \\ \prod_{k=1}^{g} \left(\frac{1}{\pi}\right)^{n_k/2} \frac{\Gamma\left(\frac{n_k+\alpha_j}{2}\right)}{\Gamma\left(\frac{\alpha_j}{2}\right)} \left(\frac{\beta_j^{\alpha_j}}{s_{jk}^{\alpha_j+n_k}}\right) \sqrt{\frac{\delta_j}{n_k+\delta_j}} \text{ if } \omega_j = 1, \end{cases} \tag{11}$$

where $s_j^2 = \beta_j^2 + \sum_{i=1}^{n}(x_{ij} - \bar{\mathrm{x}}_j)^2 + \frac{(\lambda_j - \bar{\mathrm{x}}_j)^2}{(\delta_j^{-1} + (n+\delta_j)^{-1})}$, $\bar{\mathrm{x}}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, $s_{jk}^2 = \beta_j^2 + \sum_{i=1}^{n} z_{ik}(x_{ij} - \bar{\mathrm{x}}_{jk})^2 + \frac{(\lambda_j - \bar{\mathrm{x}}_{jk})^2}{(\delta_j^{-1} + (n_k+\delta_j)^{-1})}$ and $\bar{\mathrm{x}}_{jk} = \frac{1}{n_k}\sum_{i=1}^{n} z_{ik} x_{ij}$. For j such as $\omega_j = 0$, we get $p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z}) = p(\mathbf{x}_{\bullet j}|g, \omega_j)$ since the partition does not impact the value of the integral.

3.2 The ICL criterion

The ICL criterion (Biernacki, C. and Celeux, G. and Govaert, G., 2010) carries out the model selection by focusing on the goal of clustering. It favors a model providing a partition with a strong evidence since it makes a trade-off between the model evidence and the partitioning evidence. The ICL criterion is defined by

$$\mathrm{ICL}(\boldsymbol{m}) = \ln p(\mathbf{x}, \hat{\mathbf{z}}_{\boldsymbol{m}}|\boldsymbol{m}), \tag{12}$$

where $\hat{\mathbf{z}}_{\boldsymbol{m}}$ is the partition given by the MAP rule evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}$.

When the model at hand is not the model used in the sampling scheme, the ICL criterion inherits robustness from this trade-off while the BIC criterion tends to overestimate the number of components. This phenomenon is illustrated in Section 5.1 by our numerical experiments.

Although the ICL criterion has a closed form, it requires the maximum likelihood estimates to define the partition $\hat{\mathbf{z}}_{\boldsymbol{m}}$. The time devoted to parameter estimation can become computationally prohibitive. Therefore, in this work, we introduce a new criterion avoiding this drawback.

3.3 The MICL criterion

We propose a new information criterion for model selection, named *MICL criterion* (Maximum Integrated Complete-data Likelihood). This criterion corresponds to the largest value of the integrated complete-data likelihood among all the possible partitions. Thus, the MICL criterion is defined by

$$\text{MICL}(\boldsymbol{m}) = \ln p(\mathbf{x}, \mathbf{z}^\star_{\boldsymbol{m}}|\boldsymbol{m}) \text{ with } \mathbf{z}^\star_{\boldsymbol{m}} = \text{argmax}_{\mathbf{z}} \ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{m}). \tag{13}$$

Obviously, this criterion is quite similar to the ICL criterion and inherits its main properties. In particular, it is less sensitive to model misspecification than the BIC criterion. Unlike the ICL and the BIC criteria, it does not require the maximum likelihood estimates but it implies the estimation of $\mathbf{z}^\star_{\boldsymbol{m}}$ which appears to be easily accessible (see Section 5 and Section 6). Among the models in competition, the selected model maximizes the MICL criterion and is denoted by $\boldsymbol{m}^\star$ with

$$\boldsymbol{m}^\star = \text{argmax}_{\boldsymbol{m}\in\mathcal{M}}\text{MICL}(\boldsymbol{m}). \tag{14}$$

The selected model $\boldsymbol{m}^\star$ is consistent when the number of components is known, see the proof in Appendix A. Nevertheless, like the ICL criterion, the MICL criterion lacks consistency to select the number of components if the component overlap is too strong. However, numerical experiments show its good behaviour to also select the right number of components (see Section 5.1).

4 Model selection and parameter estimation

The number of components of the competing models is usually bounded by a value $g_{\max}$. So, the space of the competing models is written as

$$\mathcal{M} = \left\{\boldsymbol{m} = (g, \boldsymbol{\omega}) :\ g \in \{1, \ldots, g_{\max}\} \text{ and } \boldsymbol{\omega} \in \{0,1\}^d\right\}. \tag{15}$$

We denote by $\mathcal{M}_g$ the restriction of $\mathcal{M}$ to the subset of the models having g components. The model $\boldsymbol{m}^\star_g$ maximizes the MICL criterion among the models belonging to $\mathcal{M}_g$. Therefore,

$$\boldsymbol{m}^\star_g = \text{argmax}_{\boldsymbol{m}\in\mathcal{M}_g}\text{MICL}(\boldsymbol{m}) \text{ with } \mathcal{M}_g = \{(g, \boldsymbol{\omega}) : \boldsymbol{\omega} \in \{0,1\}^d\}. \tag{16}$$

Thus, $\boldsymbol{m}^\star_g$ defines the best variable selection according to the MICL criterion for a fixed value of g. Obviously,

$$\boldsymbol{m}^\star = \text{argmax}_{g=1,\ldots,g_{\max}}\text{MICL}(\boldsymbol{m}^\star_g). \tag{17}$$

The estimation of $\boldsymbol{m}^\star_g$ implies to maximize the integrated complete-data likelihood on $(\boldsymbol{\omega}, \mathbf{z})$. This maximization is facilitated by the fact that $\boldsymbol{\omega}$ does not influence the definition space of the vector of component membership. Indeed, $\mathbf{z}^\star_{\boldsymbol{m}}$ is defined on the same space $\{1, \ldots, g\}^n$ for each model $\boldsymbol{m}$ in $\mathcal{M}_g$. This twofold optimization is performed by an iterative algorithm presented below. Note that this algorithm cannot be used to optimize the ICL or the BIC criterion, since $\boldsymbol{\theta}$ us not defined on the same space for each model in $\mathcal{M}_g$. We obtain $\boldsymbol{m}^\star$ by running this algorithm with g chosen from one to $g_{\max}$.

4.1 Algorithm for MICL-based model selection

The following iterative algorithm is used to find $\boldsymbol{m}^\star_g$ for any g in $\{1, \ldots, g_{\max}\}$. Starting from the initial point $(\mathbf{z}^{[0]}, \boldsymbol{m}^{[0]})$ with $\boldsymbol{m}^{[0]} \in \mathcal{M}_g$, it alternates between two optimizations of the integrated complete-data likelihood: optimization on $\mathbf{z}$, given $(\mathbf{x}, \boldsymbol{m})$, and maximization on $\boldsymbol{\omega}$ given $(\mathbf{x}, \mathbf{z})$. The algorithm is initialized as follows: first $\boldsymbol{m}^{[0]}$ is sampled from a uniform distribution on $\mathcal{M}_g$, second $\mathbf{z}^{[0]} = \hat{\mathbf{z}}_{\boldsymbol{m}^{[0]}}$ is the partition provided by a MAP rule associated to model $\boldsymbol{m}^{[0]}$ and to its maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{\boldsymbol{m}^{[0]}}$. Iteration $[r]$ of the algorithm is written as

Partition step: fix $\mathbf{z}^{[r]}$ such that

$$\ln p(\mathbf{x}, \mathbf{z}^{[r]}|\boldsymbol{m}^{[r]}) \geq \ln p(\mathbf{x}, \mathbf{z}^{[r-1]}|\boldsymbol{m}^{[r]}).$$

Model step: fix $\boldsymbol{m}^{[r+1]} = \text{argmax}_{\boldsymbol{m}\in\mathcal{M}_g} \ln p(\mathbf{x}, \mathbf{z}^{[r]}|\boldsymbol{m})$ such that

$$\boldsymbol{m}^{[r+1]} = (g, \boldsymbol{\omega}^{[r+1]}) \text{ with } \omega_j^{[r+1]} = \text{argmax}_{\omega_j\in\{0,1\}} p(\mathbf{x}_{\bullet j}|g, \omega_j, \mathbf{z}^{[r]}).$$

The partition step is performed by an iterative method. Each iteration consists in sampling uniformly an individual which is affiliated to the class maximizing the integrated complete-data likelihood while the other class memberships are unchanged.

Like an EM algorithm, the proposed algorithm converges to a local optimum of $\ln p(\mathbf{x}, \mathbf{z}|\boldsymbol{m})$. Thus, many different initializations should be used to ensure the convergence to $\boldsymbol{m}_g^\star$. However, we show that this algorithm does not suffer from the problem of local optima during our applications (see Section 6).

4.2 Maximum likelihood inference for the model maximizing the MICL criterion

When model $\boldsymbol{m}^\star = (g^\star, \boldsymbol{\omega}^\star)$ has been found, usually the estimate $\hat{\boldsymbol{\theta}}_{\boldsymbol{m}^\star}$ maximizing the likelihood function is required:

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{m}^\star} = \text{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{\boldsymbol{m}^\star}} p(\mathbf{x}|\boldsymbol{m}^\star, \boldsymbol{\theta}). \tag{18}$$

The direct optimization of the likelihood function would involve to solve equations that have no analytical solution. Instead, the parameter estimation is performed via an EM algorithm (Dempster, A.P. and Laird, N.M. and R 1977), which is often simple and efficient in the situation of missing data. This iterative algorithm alternates between two steps: the computation of the complete-data log-likelihood conditional expectation (E step) and its maximization (M step). Its iteration $[r]$ is written as:

E step: computation of the conditional probabilities

$$t_{ik}^{[r]} = \frac{\pi_k^{[r]} \prod_{j=1}^d \phi(x_{ij}|\mu_{kj}^{[r]}, \sigma_{kj}^{[r]2})}{\sum_{k'=1}^{g^\star} \pi_{k'}^{[r]} \prod_{j=1}^d \phi(x_{ij}|\mu_{k'j}^{[r]}, \sigma_{k'j}^{[r]2})}.$$

M step: maximization of the complete-data log-likelihood

$$\pi_k^{[r+1]} = \frac{t_{\bullet k}^{[r]}}{n}, \; \mu_{kj}^{[r+1]} = \begin{cases} \frac{1}{t_{\bullet k}^{[r]}} \sum_{i=1}^n t_{ik}^{[r]} x_{ij} \text{ if } \omega_j^\star = 1 \\ \frac{1}{n} \sum_{i=1}^n x_{ij} \text{ if } \omega_j^\star = 0, \end{cases}$$

$$\sigma_{kj}^{[r+1]2} = \begin{cases} \frac{1}{t_{\bullet k}^{[r]}} \sum_{i=1}^n t_{ik}^{[r]} (x_{ij} - \mu_{kj}^{[r+1]})^2 \text{ if } \omega_j^\star = 1 \\ \frac{1}{n} \sum_{i=1}^n (x_{ij} - \mu_{kj}^{[r+1]})^2 \text{ if } \omega_j^\star = 0, \end{cases}$$

where $t_{\bullet k}^{[r]} = \sum_{i=1}^n t_{ik}^{[r]}$. Note that the EM algorithm can provide the maximum *a posteriori* estimate by slightly modifying its M step (Green, P.J., 1990).

5 Numerical experiments on simulated data

Implementation of the proposed method Results of our method (indicated by MS) are provided by the R package *VarSelLCM* available on CRAN. This package performs 50 random initializations of the algorithm described in Section 4.1 to carry out the model selection. The following hyper-parameters are chosen to be fairly flat in the region where the likelihood is substantial and not much greater else-where: $\alpha_j = 1$, $\beta_j = 1$, $\lambda_j = \text{mean}(\mathbf{x}_{\bullet j})$ and $\delta_j = 0.01$.

Competing methods

- The model-based clustering method of Raftery and Dean (2006) is denoted RD in what follows. It runs using the R package *clustvarsel* (Scrucca and Raftery, 2014). Results are given by the *headlong* algorithm for the *forward* direction (denoted RD-forw) and by the *greedy* algorithm in the *backward* direction (denoted RD-back).
- The sparse K-means method of Witten and Tibshirani (2010) is not a model-based approach. It runs using the R package *sparcl* (Witten and Tibshirani, 2013) with its options by default. In what follows, this method is indicated by WT.

Simulation map First, information criteria are compared on datasets sampled from the well-specified model and on datasets sampled from a misspecified model. Second, the three competing methods of variable selection are compared. The calculations are carried out on an 8 Intel Xeon 3.40GHZ CPU machine.

5.1 Comparing model selection criteria

5.1.1 Simulated data: well-specified model

Here we compare model selection criteria when the sampling model belongs to the set of the competing models. Individuals are drawn from a bi-component Gaussian mixture model with conditional independence assumption. The first two variables are relevant to the clustering and the last two variables are not. Thus, the true model denoted by $\boldsymbol{m}^{(0)}$ is

$$\boldsymbol{m}^{(0)} = (g^{(0)}, \boldsymbol{\omega}^{(0)}) \text{ with } g^{(0)} = 2 \text{ and } \boldsymbol{\omega}^{(0)} = (1, 1, 0, 0).$$

The following parameters are used:

$$\pi_k = 0.5,\ \mu_{11} = \mu_{12} = \varepsilon,\ \mu_{21} = \mu_{22} = -\varepsilon,\ \mu_{k3} = \mu_{k4} = 0 \text{ and } \sigma_{kj} = 1.$$

The value of ε defines the class overlap. Table 1 presents the results obtained for different sample sizes and for different class overlaps. For each case, 100 samples are generated and the criteria are computed for all the possible models in $\mathcal{M}$ with $g_{\max} = 6$.

		n				
ε	criterion	50	100	200	400	800
1.26	BIC	95 (85)	100 (99)	100 (100)	100 (100)	100 (100)
	ICL	85 (72)	98 (95)	100 (97)	100 (97)	100 (99)
	MICL	86 (73)	98 (94)	100 (97)	100 (98)	100 (99)
1.05	BIC	85 (77)	99 (94)	100 (98)	100 (99)	100 (100)
	ICL	45 (42)	69 (62)	98 (94)	100 (99)	100 (100)
	MICL	50 (45)	69 (62)	99 (96)	100 (99)	100 (100)
0.85	BIC	46 (32)	73 (69)	98 (94)	100 (99)	100 (100)
	ICL	9 (7)	15 (13)	12 (9)	27 (26)	31 (31)
	MICL	10 (8)	16 (15)	13 (11)	31 (30)	35 (35)

Table 1 Results of model selection for different information criteria under the true model. In plain, percentage where the true number of components ($g^{(0)}$) has been selected. In parenthesis, percentage where the true model ($\boldsymbol{m}^{(0)}$) has been selected.

When the class overlap is not too high, all the criteria are consistent. Indeed, they asymptotically always select the true model. In such a case, the BIC criterion outperforms the other criteria when the sample size is small. When the class overlap is equal to 0.20, the BIC criterion stays consistent while the other ones select only a single class. However, we now show that the BIC criterion suffers from a lack of robustness.

5.1.2 Simulated data: misspecified model

We look at robustness of the criteria based on the integrated complete-data likelihood. Again, the first two variables contain the relevant clustering information. They follow a bi-component mixture model of uniform distributions with conditional independence assumption and equal proportions. More specifically, they are generated independently from the uniform distribution on $[\varepsilon - 1, \varepsilon + 1]$ for the first component and the uniform distribution on $[-\varepsilon - 1, -\varepsilon + 1]$ for the second. The remaining two variables are irrelevant variables that are independent of the clustering variables and follow two independent standard Gaussian distributions. For each case, 100 samples are generated and the criteria are computed for all the possible models in $\mathcal{M}$ with $g_{\max} = 6$. Table 2 summarizes the selection results for each criterion.

Results show that the BIC criterion is not useful to select the number of components. Indeed, it overestimates the number of classes to better fit the data since the sampling model does not belong to the set of the competing models. The other criteria show considerably better performance since they select the true number of classes and the true $\boldsymbol{\omega}$. It appears that they are more robust than the BIC criterion to the misspecification of the model at hand.

To conclude, the ICL and the MICL criteria obtain good results for model selection when the class overlap is not too strong. Moreover, they are more robust to model misspecification than the BIC criterion. Since the MICL criterion does not require maximum likelihood inference for all of the competing models, it is preferable to the ICL criterion for carrying out model selection.

		n				
ε	criterion	50	100	200	400	800
1.26	BIC	79 (75)	80 (79)	48 (48)	0 (0)	0 (0)
	ICL	100 (96)	100 (98)	100 (100)	100 (99)	97 (97)
	MICL	100 (95)	100 (98)	100 (100)	100 (99)	96 (96)
1.05	BIC	86 (83)	83 (80)	49 (48)	4 (4)	0 (0)
	ICL	100 (91)	100 (95)	100 (98)	99 (98)	99 (98)
	MICL	100 (92)	100 (99)	100 (98)	98 (98)	99 (98)
0.85	BIC	80 (78)	72 (71)	36 (36)	0 (0)	0 (0)
	ICL	97 (87)	100 (96)	100 (98)	99 (97)	97 (97)
	MICL	97 (92)	100 (98)	100 (98)	99 (97)	97 (97)

Table 2 Results of model selection for different information criteria under the non-Gaussian model. In plain, percentage where the true number of components ($g = 2$) has been selected. In parenthesis, percentage where the true number of classes and the true partitioning of the variable ($\boldsymbol{\omega} = (1, 1, 0, 0)$) have been selected.

5.2 Comparing methods on simulated data

Data are drawn from a tri-component Gaussian mixture model assuming conditional independence and equal proportions. The first r variables are relevant while the last $d - r$ variables are irrelevant since they follow standard Gaussian distributions. Under component k, the first r variables follow a spherical Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_k; \boldsymbol{I})$ with $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = (\varepsilon, \ldots, \varepsilon) \in \mathbb{R}^r$ and $\boldsymbol{\mu}_3 = \mathbf{0}_r$. The three competing methods are compared on five different scenarios described in Table 3.

	n	r	d	ε
Scenario 1	30	5	25	0.6
Scenario 2	30	5	25	1.7
Scenario 3	300	5	25	1.7
Scenario 4	300	5	100	1.7
Scenario 5	300	50	500	1.7

Table 3 The five scenarios used for the method comparisons.

For each scenario, 25 samples of size n are generated and the analysis is performed with $g = 3$. Results are presented in Table 4. Note that the RD method is run only on the smaller scenarios for computational reasons.

Scenario	Method	NRV	RRR	RIR	ARI	Time
1	MS	1.28	0.08	0.95	0.01	0.42
	WT	8.28	0.38	0.68	0.06	1.78
	RD-forw	3.96	0.15	0.84	0.06	1.15
	RD-back	11.00	0.42	0.55	0.03	1.30
2	MS	5.40	1.00	0.98	0.66	0.47
	WT	12.88	0.96	0.59	0.59	1.70
	RD-forw	3.28	0.15	0.87	0.13	0.97
	RD-back	10.64	0.58	0.61	0.37	49.01
3	MS	5.00	1.00	1.00	0.86	16.21
	WT	25.00	1.00	0.00	0.87	7.97
	RD-forw	5.52	0.96	0.96	0.82	6.42
	RD-back	5.64	1.00	0.97	0.86	30.65
4	MS	5.00	1.00	1.00	0.88	73.36
	WT	100.00	1.00	0.00	0.89	21.22
	RD-forw	6.96	0.92	0.97	0.81	38.05
	MS	50.04	1.00	1.00	1.00	48.37
5	WT	500	1.00	0.00	1.00	83.49

Table 4 Comparing variable selection methods on simulated data. Means of the numbers of relevant variables (NRV), the right relevant rates (RSR), the right irrelevant rates (RIR), the Adjusted Rand Indices (ARI) and the computing times in second (Time).

When the sample size is small and when the component overlap is high (Scenario 1), all methods obtain poor results. However, when the component separation increases (Scenario 2), MS method leads to a better variable selection (NRV, RRR, RIR) which involves a better partitioning accuracy (ARI). Indeed, WT selects some variables which are not discriminative and which damage the partitioning accuracy. Both directions of RD lead to only average results.

When the sample size increases (Scenarios 3, 4 and 5), the MS results for the variable selections are ameliorated and the true model is almost always found. In this context, WT claims that all the variables are relevant to the clustering. Since the sample size is not too small, its partitioning accuracy is not damaged but the model interpretation is strongly harder since all the variables should be used for characterizing the classes. In this context, the RD results are good when the number of variables is small, but they are damaged when many variables are observed. Moreover, when many variables are observed (Scenarios 5), the multiple calls of EM algorithm for the model comparisons prevent the using of this method computational reasons.

During these experiments, the algorithm assessing $\boldsymbol{m}^\star$ does not suffer from strong problems of local optima except for Scenario 1. Indeed, Table 5 presents statistics of the occurrence where the couple $(\mathbf{z}^\star_{\boldsymbol{m}^\star}, \boldsymbol{m}^\star)$ has been found by the algorithm for the 50 random initializations.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
Mean	13	17	31	19	45
Min	1	4	9	8	10
Max	35	31	45	28	48

Table 5 Mean and minimal occurrence where the couple $(\mathbf{z}^\star_{\boldsymbol{m}^\star}, \boldsymbol{m}^\star)$ has been found by the algorithm for the 50 random initializations for the 25 generated data.

6 Numerical experiments on benchmark data

We now compare the competing methods on real datasets in which the correct number of groups is known. Theses data sets are presented in Table 6.

Name	d	n	$g^{(0)}$	Reference	R package/website
banknote	6	200	2	Flury, B. and Riedwyl, H. (1988)	VarSelLCM
coffee	12	43	2	Streuli (1973)	ppgm
wine	13	178	3	Forina and al (1991)	UCI
cancer	30	569	2	Street et al. (1993)	UCI
golub	3051	83	2	Golub and al. (1999)	multtest

Table 6 Information about the benchmark data sets.

Our study is divided in two parts: first the three competing methods are compared by assuming the component number known, second the model-based methods are compared by assuming the component number unknown. Because RD-back is very slow, we perform the RD method only with the forward approach. The experimental conditions are similar to those described in the previous section. Since the sparse K-means method does not provide an automatic procedure to select the number of groups, this method is not not used in the second part.

6.1 Component number known

Table 7 presents the results obtained when the number of components is known.

The first three data sets are the less challenging since the number of variables is moderate. However, MS provides an easier interpretable model than WT (less relevant variables) which provides an identical partitioning accuracy. Surprisingly, MS leads to a model having a better BIC value than RD while this latter method aims to maximizing this criterion. This phenomenon illustrates the sub-optimality problem

Data	NRV			ARI			Time			BIC	
	MS	WT	RD	MS	WT	RD	MS	WT	RD	MS	RD
banknote	5	6	4	0.96	0.96	0.98	2.1	2.3	1.6	-968	-1009
coffee	5	12	5	1.00	1.00	1.00	0.1	0.9	0.7	-522	-555
wine	11	13	5	0.87	0.85	0.73	5.4	2.4	1.5	-3538	-3769
cancer	15	30	17	0.75	0.70	0.65	50	17	62	2189	2569
golub	553	3051	10	0.79	0.11	0.00	34	54	258	-90348	-95255

Table 7 Results obtained when the component number is known: number of relevant variables (NRV), adjusted rand index computed on the selected model (ARI), computing time in seconds (Time) and BIC criterion value.

of the RD procedure. Moreover, note that, by selecting only five variables on *wine* data set, RD provides a less relevant partition.

For the larger data sets, RD obtains better results. Indeed, it is more easily interpretable and it provides a more accurate partition. For instance, on *cancer* data set, MS obtains a better ARI than both other methods while it selects less variables. Concerning *golub* data set, WT relates all the variables while RD selects only ten variables. These methods also obtain a worse partition than MS. Finally, note that MS obtains a better value of the BIC criterion than RD on *golub* data set.

Table 8 shows that the couple $(\mathbf{z}^\star_{\boldsymbol{m}^\star}, \boldsymbol{m}^\star)$ is easily accessible by the proposed algorithm. Thus, problems due to multiple local optima do not occur on these data sets. Moreover, it indicates the values of the ICL and MICL criteria related to the model selected by MS. Even if the MICL value is always larger (or equal) than the ICL value, this difference is often small. Thus, the partitions $\hat{\mathbf{z}}_{\boldsymbol{m}^\star}$ and $\mathbf{z}^\star_{\boldsymbol{m}^\star}$ are often quite similar.

	banknote	coffee	wine	cancer	golub
Occurrence	48	27	11	48	6
MICL	-1009.2	-644.1	-3715.7	-7963.5	-103858.8
ICL	-1009.2	-644.1	-3715.8	-8064.1	-103858.8

Table 8 Ocurrence where the couple $(\mathbf{z}^\star_{\boldsymbol{m}^\star}, \boldsymbol{m}^\star)$ has been found by the algorithm for the 50 random initializations and values of the information criteria for model $\boldsymbol{m}^\star$.

6.2 Component number unknown

Table 9 presents the results obtained when the number of components is unknown by setting $g_{\max} = 6$.

Data	$\hat{g}$		NRV		ARI		BIC	
	MS	RD	MS	RD	MS	RD	MS	RD
banknote	3	5	6	4	0.61	0.40	-926	-978
coffee	2	3	5	6	1.00	0.38	-522	-521
wine	4	4	11	6	0.67	0.72	-3502	-3739
cancer	6	6	13	15	0.21	0.23	4192	4861
golub	2	6	553	8	0.79	0.00	-90348	-95098

Table 9 Results obtained when the component number is unknown: number of components ($\hat{g}$), number of relevant variables (NRV), adjusted rand index computed on the selected model (ARI) and BIC criterion value.

On these data sets, RD selects more components than MS. Thus, its interpretation becomes more complex even if RD can select less variables. The previous remarks are validated by this application. Indeed, we observe than MS can provide a model with a better BIC value. Moreover, when the number of variables increases, the results of RD are damaged due to the lack of optimality. This is shown particularly by *golub* data set.

7 Discussion

We have proposed a new information criterion to carry out model selection of a finite mixture model. This criterion can be used for selecting the relevant variables for model-based clustering in Gaussian mixture settings assuming conditional independence. In such a case, the criterion has a closed form and the model maximizing it is accessible by an algorithm of alternated optimization. Its originality consists in allowing a model selection procedure which does not require the maximum likelihood estimate.

The criterion can be easily used when the model at hand is a mixture of distributions belonging to an exponential family. Indeed, in such cases, the closed form is preserved. Thus, the MICL criterion can carry out variable selection in a cluster analysis of categorical or mixed datasets by using the model of Celeux and Govaert (1991) and of Moustaki, I. and Papageorgiou, I. (2005) respectively. The application of the proposed method to the categorical data appears to be especially pertinent since conditional independence assumption is often assumed for such data, and since Jeffreys non informative prior distributions are available.

If the conditional independence assumption is relaxed, the algorithm used for model selection should be modified. Then, its model step is not explicit but it can be achieved by a MCMC method.

We have compared our method with two standard procedures of variable selection in cluster analysis. It was shown that the proposed method outperforms both other ones for the task of variable selection. It results in a better partitioning accuracy. In a moderate computing time, the proposed method can manage datasets with a large number of variables and a relatively large number of individuals. However, the procedure of model selection is time-consuming if a huge number of individuals is observed. In such a case, the optimization of the model selection procedure is an issue which calls for further improvements.

Finally, this approach could be extend to perform a more elaborated variable selection. Indeed, by using the approach of Maugis, C. and Celeux, G. and Martin-Magniette, M. (2009).

The R package *VarSelLCM* implementing the proposed method is downloadable on CRAN.

Acknowledgement

The authors are grateful to Gilles Celeux and Jean-Michel Marin for their leading comments.

References

Biernacki, C. and Celeux, G. and Govaert, G. (2010). Exact and Monte Carlo calculations of integrated likelihoods for the latent class model. *Journal of Statistical Planning and Inference*, 140(11):2991–3002.

Celeux, G. and Govaert, G. (1991). Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8(2):157–176.

Dempster, A.P. and Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics: A practical approach*. London: Chapman and Hall.

Forina, M. and al (1991). PARVUS An Extendible Package for Data Exploration, Classification and Correlation. *Institute of Pharmaceutical and Food Analysis and Technologies, Genoa, Italy*.

Friedman, J.H. and Meulman, J.J. (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(4):815–849.

Friel, N. and Wyse, J. (2012). Estimating the evidence–a review. *Statistica Neerlandica*, 66(3):288–308.

Golub, T. and al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.

Govaert, G. (2009). *Data Analysis*. ISTE Wiley.

Green, P.J. (1990). On use of the EM for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):443–452.

Hand, D. and Keming, Y. (2001). Idiot's Bayes, not so stupid after all? *International statistical review*, 69(3):385–398.

Maugis, C. and Celeux, G. and Martin-Magniette, M. (2009). Variable selection for clustering with Gaussian mixture models. *Biometrics*, 65(3):701–709.

Moustaki, I. and Papageorgiou, I. (2005). Latent class models for mixed variables with applications in Archaeometry. *Computational Statistics and Data Analysis*, 48(3):659 – 675.

Pan, W. and Shen, X. (2007). Penalized Model-Based Clustering with Application to Variable Selection. *Journal of Machine Learning Research*, 8:1145–1164.

Raftery, A. and Dean, N. (2006). Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, 101(473):168–178.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Scrucca, L. and Raftery, A. (2014). clustvarsel: A Package Implementing Variable Selection for Model-based Clustering in R. *(submitted to) Journal of Statistical Software*.

Street, W., Wolberg, W., and Mangasarian, O. (1993). Nuclear feature extraction for breast tumor diagnosis. *IST/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, 1905:861–870.

Streuli, H. (1973). Der heutige stand der kaffeechemie. *In Association Scientifique International du Cafe, 6th International Colloquium on Coffee Chemisrty*, pages 61–72.

Tadesse, M.G. and Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617.

Witten, D. and Tibshirani, R. (2010). A Framework for Feature Selection in Clustering. *Journal of the American Statistical Association*, 105(490):713–726.

Witten, D. and Tibshirani, R. (2013). *sparcl: Perform sparse hierarchical clustering and sparse k-means clustering*. R package version 1.0.3.

Appendices

A Consistency of the MICL criterion

This section is devoted to the proof of consistency of our MICL criterion with a fixed number of components. The first part deals with non-nested models and requires a *biais-entropy* compensation assumption. The second part covers the nested models, *i.e*, when the competing model contains the true model. In what follows, we consider the true model $\boldsymbol{m}^{(0)} = \big(g^{(0)}, \boldsymbol{\omega}^{(0)}\big)$, its set of relevant variables is $\Omega^{(0)} = \Big\{j : \omega_j^{(0)} = 1\Big\}$ and the parameter is $\boldsymbol{\theta}^{(0)}$.

Case of non-nested model We need to introduce the entropy notation given by

$$\xi\big(\boldsymbol{\theta}; \mathbf{z}, \boldsymbol{m}\big) = \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} \ln \tau_{ik}\big(\boldsymbol{\theta} \mid \boldsymbol{m}\big),$$

where $\tau_{ik}\big(\boldsymbol{\theta} \mid \boldsymbol{m}\big) = \dfrac{\pi_k \phi\big(\boldsymbol{x}_i \mid \theta_k, \boldsymbol{m}\big)}{\sum_h^g \pi_h \phi\big(\boldsymbol{x}_i \mid \theta_h, \boldsymbol{m}\big)}$.

Proposition 1 *Assume that $\boldsymbol{m}^{(1)}$ is a model such that $\boldsymbol{m}^{(0)}$ is a non-nested within $\boldsymbol{m}^{(1)}$. Assume that*

$$-\mathbb{E}\left[\ln \frac{\sum_{k=1}^{g^{(0)}} \pi_k \prod_{j=1}^{d} \phi\big(x_{1j} \mid \mu_{kj}^{(0)}, \sigma_{kj}^{(0)2}\big) \mathbb{1}_{G_k^{(0)}}\big(\boldsymbol{x}_1\big)}{p\big(\boldsymbol{x}_1 \mid \boldsymbol{\theta}^{(0)}, \boldsymbol{m}^{(0)}\big)}\right] \leq \mathbf{KL}\Big[\boldsymbol{m}^{(0)} || \boldsymbol{m}^{(1)}\Big], \tag{19}$$

where $\mathbf{KL}\Big[\boldsymbol{m}^{(0)} || \boldsymbol{m}^{(1)}\Big]$ is the Kullback-Leibler divergence of $p\big(\cdot \mid \boldsymbol{\theta}^{(0)}, \boldsymbol{m}^{(0)}\big)$ from $p\big(\cdot \mid \boldsymbol{\theta}^{(1)}, \boldsymbol{m}^{(1)}\big)$ and

$$G_k^{(0)} = \left\{ x \in \mathbb{R}^d : k = \underset{1 \leq h \leq g^{(0)}}{argmax}\ \pi_h \prod_{j=1}^{d} \phi\big(x_{1j} \mid \mu_{hj}^{(0)}, \sigma_{hj}^{2(0)}\big) \right\}.$$

When $n \to \infty$, we have

$$\mathbb{P}\Big(MICL\big(\boldsymbol{m}^{(1)}\big) > MICL\big(\boldsymbol{m}^{(0)}\big) \Big) \longrightarrow 0.$$

Proof For any model $\boldsymbol{m}$, we have the following inequalities,

$$\mathrm{ICL}\big(\boldsymbol{m}\big) \leq \mathrm{MICL}\big(\boldsymbol{m}\big) \leq \ln p\big(\mathbf{x} \mid \boldsymbol{m}\big).$$

It follows,

$$\mathbb{P}\Big\{\mathrm{MICL}\big(\boldsymbol{m}^{(1)}\big) - \mathrm{MICL}\big(\boldsymbol{m}^{(0)}\big) > 0\Big\} \leq \mathbb{P}\Big\{\ln p\big(\mathbf{x} \mid \boldsymbol{m}^{(1)}\big) - \mathrm{ICL}\big(\boldsymbol{m}^{(0)}\big) > 0\Big\}.$$

Now set $\Delta\nu = \nu^{(1)} - \nu^{(0)}$ where $\nu^{(1)}$ and $\nu^{(0)}$ are the numbers of free parameters in the models $\boldsymbol{m}^{(1)}$ and $\boldsymbol{m}^{(0)}$ respectively. Using Laplace's approximation, we have

$$\mathrm{ICL}\big(\boldsymbol{m}^{(0)}\big) = \ln p\left(\mathbf{x} \mid \widehat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{m}^{(0)}\right) + \xi\left(\widehat{\boldsymbol{\theta}}^{(0)}; \widehat{\mathbf{z}}^{(0)}, \boldsymbol{m}^{(0)}\right) - \frac{\nu^{(0)}}{2}\ln n + \mathcal{O}_p(1),$$

where $\widehat{\boldsymbol{\theta}}^{(0)}$ and $\widehat{\mathbf{z}}^{(0)}$ are respectively the MLE and the partition given by the corresponding MAP rule. In the same way, we have

$$\ln p\big(\mathbf{x} \mid \boldsymbol{m}^{(1)}\big) = \ln p\big(\mathbf{x} \mid \widehat{\boldsymbol{\theta}}^{(1)}, \boldsymbol{m}^{(1)}\big) - \frac{\nu^{(1)}}{2}\ln n + \mathcal{O}_p(1),$$

where $\widehat{\boldsymbol{\theta}}^{(1)}$ is the MLE of $\boldsymbol{\theta}^{(1)}$. Note that

$$\ln p\big(\mathbf{x} \mid \boldsymbol{m}^{(1)}\big) - \mathrm{ICL}\big(\boldsymbol{m}^{(0)}\big) = \frac{A_n}{2} + nB_n - \frac{\Delta\nu}{2}\ln n + \mathcal{O}_p(1),$$

where

$$A_n = 2\ln\frac{p\left(\mathbf{x} \mid \widehat{\boldsymbol{\theta}}^{(1)}, \boldsymbol{m}^{(1)}\right)}{p\left(\mathbf{x} \mid \boldsymbol{\theta}^{(1)}, \boldsymbol{m}^{(1)}\right)} - 2\ln\frac{p\left(\mathbf{x} \mid \widehat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{m}^{(0)}\right)}{p\left(\mathbf{x} \mid \boldsymbol{\theta}^{(0)}, \boldsymbol{m}^{(0)}\right)},$$

and

$$B_n = \frac{1}{n}\ln\frac{p\left(\mathbf{x} \mid \boldsymbol{\theta}^{(1)}, \boldsymbol{m}^{(1)}\right)}{p\left(\mathbf{x} \mid \boldsymbol{\theta}^{(0)}, \boldsymbol{m}^{(0)}\right)} - \frac{1}{n}\xi\left(\widehat{\boldsymbol{\theta}}^{(0)}; \widehat{\mathbf{z}}^{(0)}, \boldsymbol{m}^{(0)}\right).$$

When $n \to \infty$, we have $A_n \to \chi^2_{\Delta\nu}$ in distribution and B_n tends to

$$-\mathbf{KL}\Big[\boldsymbol{m}^{(0)} || \boldsymbol{m}^{(1)}\Big] - \mathbb{E}\left[\ln\frac{\sum_{k=1}^{g^{(0)}}\pi_k\prod_{j=1}^{d}\phi\big(x_{1j} \mid \mu_{kj}^{(0)}, \sigma_{kj}^{(0)2}\big)\mathbb{1}_{G_k^{(0)}}\big(\boldsymbol{x}_1\big)}{p\big(\boldsymbol{x}_1 \mid \boldsymbol{\theta}^{(0)}, \boldsymbol{m}^{(0)}\big)}\right]$$

in probability. Thus, under the assumption (19), MICL is consistent since when $n \to \infty$, we have

$$\begin{aligned}\mathbb{P}\Big\{\mathrm{MICL}\big(\boldsymbol{m}^{(1)}\big) - \mathrm{MICL}\big(\boldsymbol{m}^{(0)}\big) > 0\Big\} &\leq \mathbb{P}\Big[A_n + \mathcal{O}_p(1) > \Delta\nu\ln n\Big] + \mathbb{P}\Big[B_n > 0\Big] \\ &\longrightarrow 0.\end{aligned}$$

Case of nested model Recall that $\mathrm{MICL}\big(\boldsymbol{m}^{(0)}\big) = \ln p\big(\mathbf{x}, \mathbf{z}^{(0)} \mid \boldsymbol{m}^{(0)}\big)$, where $\mathbf{z}^{(0)} = \underset{\mathbf{z}}{\mathrm{argmax}}\ln p\big(\mathbf{x}, \mathbf{z} \mid \boldsymbol{m}^{(0)}\big)$. We have

$$\mathbf{z}^{(0)} = \underset{\mathbf{z}}{\mathrm{argmax}}\Big\{\ln p\left(\mathbf{z} \mid g^{(0)}\right) + \sum_{j\in\Omega_0}\ln p\big(\mathbf{x}_{\bullet j} \mid \omega_j^{(0)}, g^{(0)}, \mathbf{z}\big)\Big\},$$

where $\Omega_0 = \Big\{j : \omega_j^{(0)} = 1\Big\}$. Let $\boldsymbol{m}^{(1)} = \big(g^{(0)}, \Omega_1\big)$ where $\Omega_1 = \Omega_0 \cup \Omega_{01}$ and $\Omega_{01} = \Big\{j : \omega_j^{(1)} = 1, \omega_j^{(0)} = 0\Big\}$. Then, in the same way, we have $\mathrm{MICL}\big(\boldsymbol{m}^{(1)}\big) = \ln p\big(\mathbf{x}, \mathbf{z}^{(1)} \mid \boldsymbol{m}^{(1)}\big)$, where

$$\mathbf{z}^{(1)} = \underset{\mathbf{z}}{\mathrm{argmax}}\left[\ln p\left(\mathbf{z} \mid g^{(0)}\right) + \sum_{j\in\Omega_1}\ln p\left(\mathbf{x}_{\bullet j} \mid \omega_j^{(1)}, g^{(0)}, \mathbf{z}\right)\right].$$

Let $j \in \Omega_{01}$, Laplace's approximation gives us,

$$\ln p\left(\mathbf{x}_{\bullet j} \mid \omega_j^{(1)}, g^{(0)}, \mathbf{z}\right) = \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} \ln \phi\left(x_{ij} \mid \tilde{\mu}_{kj}^{(1)}, \tilde{\sigma}_{kj}^{(1)2}\right) - g^{(0)} \ln n + \mathcal{O}_p(1),$$

where

$$\left(\tilde{\mu}_{kj}^{(1)}, \tilde{\sigma}_{kj}^{(1)2}\right) = \underset{\mu_{kj}^{(1)}, \sigma_{kj}^{(1)2}}{\operatorname{argmax}} \sum_{i=1}^{n} z_{ik} \ln \phi\left(x_{ij} \mid \mu_{kj}^{(1)}, \sigma_{kj}^{(1)2}\right).$$

Proposition 2 *Assume that $\boldsymbol{m}^{(1)}$ is a model such that $g^{(1)} = g^{(0)}$ and $\Omega_1 = \Omega_0 \cup \Omega_{01}$ where $\Omega_{01} \neq \emptyset$, i.e, the model $\boldsymbol{m}^{(0)}$ is nested within the model $\boldsymbol{m}^{(1)}$ with the same number of components. When $n \to \infty$,*

$$\mathbb{P}\Big(\mathit{MICL}\big(\boldsymbol{m}^{(1)}\big) > \mathit{MICL}\big(\boldsymbol{m}^{(0)}\big)\Big) \longrightarrow 0.$$

Proof We have

$$\mathbb{P}\Big\{\mathrm{MICL}\big(\boldsymbol{m}^{(1)}\big) > \mathrm{MICL}\big(\boldsymbol{m}^{(0)}\big)\Big\} \leq \mathbb{P}\left\{\sum_{j\in\Omega_{01}} \ln \frac{p\big(\mathbf{x}_{\bullet j} \mid \omega_j^{(1)}, g^{(0)}, \mathbf{z}^{(1)}\big)}{p\big(\mathbf{x}_{\bullet j} \mid \omega_j^{(0)}, g^{(0)}, \mathbf{z}^{(0)}\big)} > 0\right\},$$

And for each $j \in \Omega_{01}$, when $n \to \infty$

$$2\sum_{i=1}^{n}\sum_{k=1}^{g^{(0)}} z_{ik}^{(1)} \ln \frac{\phi\left(x_{ij} \mid \tilde{\mu}_{kj}^{(1)}, \tilde{\sigma}_{kj}^{(1)2}\right)}{\phi\left(x_{ij} \mid \mu_{1j}^{(0)}, \sigma_{1j}^{(0)2}\right)} \longrightarrow \chi^2_{2g^{(0)}} \quad \text{in distribution.}$$

We have

$$\begin{aligned}\mathbb{P}\Big(\sum_{j\in\Omega_{01}} \ln \frac{p\big(\mathbf{x}_{\bullet j} \mid \omega_j^{(1)}, g^{(0)}, \mathbf{z}^{(1)}\big)}{p\big(\mathbf{x}_{\bullet j} \mid \omega_j^{(0)}, g^{(0)}, \mathbf{z}^{(0)}\big)} > 0\Big) &= \mathbb{P}\Big(\chi^2_{2(g^{(0)}-1)} - 2(g^{(0)}-1)\ln n > 0\Big) \\ &\longrightarrow 0 \quad \text{by Chebyshev's inequality.}\end{aligned}$$